Eine lineare Funktion kann immer in die Form $f(x) = mx + b$ mit $m, b \in \mathbb{R}$ gebracht werden. Bei vielen Experimenten ergeben sich lineare Zusammenhänge, trotzdem liegen die Messwerte in der Regel nie exakt auf einer Geraden. Man bestimmt dann eine Ausgleichsgerade, die zu den Messwerten möglichst gut passt.

Grundlage arithmetisches Mittel:
Das arithmetische Mittel von den Werten $x_1, x_2, \ldots, x_n$ ist definiert als

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

Beispiele: Das arithmetische Mittel der Werte $\{2, 3, 5, 14\}$ ist

$$\bar{x} = \frac{2 + 3 + 5 + 14}{4} = \frac{24}{4} = 6$$

Das arithmetische Mittel der Werte $\{-3, 11, -5, -7, 20\}$ ist

$$\bar{x} = \frac{-3 + 11 - 5 - 7 + 20}{5} = \frac{16}{5} = 3.2$$

Regressionsgerade:
Gegeben sind $n \in \mathbb{N}$ Messpunkte im $\mathbb{R}^2$, d.h. $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. Sei $P = x_1y_1 + x_2y_2 + \cdots x_ny_n$ und $S = x_1^2 + x_2^2 + \cdots x_n^2$. Die Regressionsgerade $r(x)$ hat die Gleichung $r(x) = mx + b$ mit

$$m = \frac{P - n\bar{x}\bar{y}}{S - n\bar{x}^2} = \frac{x_1y_1 + x_2y_2 + \cdots x_ny_n - n\bar{x}\bar{y}}{x_1^2 + x_2^2 + \cdots x_n^2 - n\bar{x}^2} \qquad b = \bar{y} - m\bar{x} \tag{1}$$

Beispiel: Gegeben sind die Punkte $(2, 4), (1, 5), (3, 6)$.

Da es drei Punkte sind ist $n = 3$. Berechne das arithmetische Mittel der $x-$Werte und der $y-$Werte:

$$\bar{x} = \frac{2 + 1 + 3}{3} = 2 \qquad \bar{y} = \frac{4 + 5 + 6}{3} = 5$$

Berechne die Summe $P = x_1y_1 + x_2y_2 + \cdots x_ny_n$

$$2 \cdot 4 + 1 \cdot 5 + 3 \cdot 6 = 31$$

Berechne die Summe $S = x_1^2 + x_2^2 + \cdots x_n^2$:

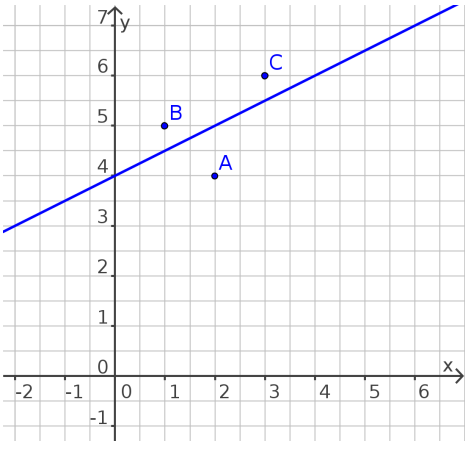
$$2^2 + 1^2 + 3^2 = 14$$

Jetzt können m und b aus Gleichung (1) berechnet werden:

$$m = \frac{31 - 3 \cdot 2 \cdot 5}{14 - 3 \cdot 2^2} = \frac{1}{2} = 0.5 \qquad b = 5 - \frac{1}{2} \cdot 2 = 4$$

Die Regressionsgerade ist:

$$r(x) = \frac{1}{2}x + 4$$



Formel für die Regressionsgerade:
Gesucht sind die Parameter m und b, so dass die Gerade $r(x) = mx+b$ möglichst gut zu den Punkten $(x_1, y_1), \ldots, (x_n, y_n)$ passt. Es ist naheliegend, dass der geometrische Schwerpunkt $(\bar{x}, \bar{y})$ der Punkteverteilung auf der Regressionsgerade liegen sollte. Daher gilt $\bar{y} = m\bar{x} + b$ und damit

$$b = \bar{y} - m\bar{x}$$

Für die Regressionsgerade gilt also:

$$r(x) = mx + \bar{y} - m\bar{x} = m(x - \bar{x}) + \bar{y}$$

Die Mittelwerte $\bar{x}$, $\bar{y}$ sind einfach zu errechnen. Es fehlt also nur noch der Parameter m. Betrachte dazu einen Punkt (x_k, y_k). An der Stelle x_k ist der Funktionswert der Regressionsgerade

$$r(x_k) = m(x_k - \bar{x}) + \bar{y}$$

und die **vertikale** Distanz d_k zum Punkt (x_k, y_k) ist:

$$d_k = y_k - r(x_k) = y_k - [m(x_k - \bar{x}) + \bar{y}] = y_k - m(x_k - \bar{x}) - \bar{y} = y_k - \bar{y} - m(x_k - \bar{x})$$

Das Quadrat hiervon ist

$$d_k^2 = \{y_k - \bar{y} - m(x_k - \bar{x})\}^2 = m^2(x_k - \bar{x})^2 - 2(x_k - \bar{x})(y_k - \bar{y})m + (y_k - \bar{y})^2$$

und es sei $Q(m)$ die Summe dieser Quadrate:

$$\begin{aligned} Q(m) &= d_1^2 + \cdots + d_n^2 \\ &= m^2(x_1 - \bar{x})^2 - 2(x_1 - \bar{x})(y_1 - \bar{y})m + (y_1 - \bar{y})^2 + \cdots + m^2(x_n - \bar{x})^2 - 2(x_n - \bar{x})(y_n - \bar{y})m + (y_n - \bar{y})^2 \\ &= \left[(x_1 - \bar{x})^2 + \cdots (x_n - \bar{x})^2\right] m^2 - 2[(x_1 - \bar{x})(y_1 - \bar{y}) + \cdots (x_n - \bar{x})(y_n - \bar{y})]m + (y_1 - \bar{y})^2 + \cdots (y_n - \bar{y})^2 \end{aligned}$$

Offensichtlich ist $Q(m)$ eine Parabel der Form

$$Q(m) = a_2 m^2 + a_1 m + a_0$$

wobei

$$\begin{aligned} a_2 &= (x_1 - \bar{x})^2 + \cdots (x_n - \bar{x})^2 \\ a_1 &= -2[(x_1 - \bar{x})(y_1 - \bar{y}) + \cdots (x_n - \bar{x})(y_n - \bar{y})] \\ a_0 &= (y_1 - \bar{y})^2 + \cdots (y_n - \bar{y})^2. \end{aligned}$$

Wegen $a_2 > 0$ ist die Parabel $Q(m)$ nach oben geöffnet, besitzt also ein Minium. Aus

$$0 = \frac{dQ}{dm} = 2a_2 m + a_1$$

folgt

$$m = -\frac{a_1}{2a_2} \tag{2}$$

Der Parameter a_2 lässt sich wie folgt vereinfachen:

$$a_2 = x_1^2 - 2x_1\bar{x} + \bar{x}^2 + \cdots + x_n^2 - 2x_n\bar{x} + \bar{x}^2 = \underbrace{x_1^2 \cdots + x_n^2}_{=S} - 2\left(\underbrace{x_1 + \cdots + x_n}_{=n\bar{x}}\right)\bar{x} + n\bar{x}^2 = S - n\bar{x}^2$$

Analog findet man für den Parameter a_1:

$$\begin{aligned} a_1 &= -2[x_1y_1 - x_1\bar{y} - y_1\bar{x} + \bar{x}\bar{y} + \cdots + x_ny_n - x_n\bar{y} - y_n\bar{x} + \bar{x}\bar{y}] \\ &= -2\left[\underbrace{x_1y_1 + \cdots + x_ny_n}_{=P} - \left(\underbrace{x_1 + \cdots + x_n}_{=n\bar{x}}\right)\bar{y} - \left(\underbrace{y_1 + \cdots + y_n}_{=n\bar{y}}\right)\bar{x} + n\bar{x}\bar{y}\right] = -2[P - n\bar{x}\bar{y}] \end{aligned}$$

Die Steigung der Regressionsgerade findet man, indem man die Summe der quadratischen Abweichungen $Q(m)$ minimiert. Das Minimum von $Q(m)$ befindet sich nach Gleichung (2) an der Stelle

$$m = -\frac{a_1}{2a_2} = -\frac{-2[P - n\bar{x}\bar{y}]}{2(S - n\bar{x}^2)} = \frac{P - n\bar{x}\bar{y}}{S - n\bar{x}^2}$$